Toward a bilingual lexical database on connectives: Exploiting a German/Italian parallel corpus

Peter Bourgonje, Yulia Grishina, Manfred Stede
Applied Computational Linguistics
University of Potsdam / Germany
{bourgonje,grishina,stede}@uni-potsdam.de

Abstract

English. We report on experiments to validate and extend two language-specific connective databases (German and Italian) using a word-aligned corpus. This is a first step toward constructing a bilingual lexicon on connectives that are connected via their discourse senses.

Italiano. *Presentiamo una serie di esperimenti per validare ed estendere due database dei connettivi, che sonospecifici per la lingua italiana e per quella tedesca. Abbiamo utilizzato un corpus parallelo allineato a livello della parola. Si tratta di un primo passo verso la costruzione di un lessico bilingue dei connettivi che sono collegati attraverso i loro sensi del discorso.*

1 Introduction

An important part of discourse processing deals with uncovering coherence relations that hold between individual, “elementary” units of a text. The lexical items that can signal such a relation are referred to as *discourse connectives*, and examples of these relations, also called the connectives’ senses, are *contrast* (e.g., ‘but’), *elaboration* (e.g., ‘in particular’), or *cause* (e.g., ‘therefore’). Notice, however, that relations need not always be signalled in text, if the context or world knowledge is sufficient for the reader to infer it, as (1)-(4) demonstrate:

(1) We should hurry, because it’s late.

(2) We should hurry. It’s late.

(3) The red pen costs $2, while the blue one is $2.50.

(4) The red pen costs $2; the blue one is $2.50.

On the other hand, example (6) is a perfectly grammatical sentence but the meaning is different from (5), so for this case of a Concession relation, the connective is in fact indispensable.

(5) Although it is late, we don’t need to hurry.

(6) It is late; we don’t need to hurry.

Recognizing these relations, which can hold within a sentence, between two sentences, or between larger spans of text, is a central task for uncovering the structure of a text, as it has been studied in theories like Rhetorical Structure Theory (Mann and Thompson, 1988) or Segmented Discourse Representation Theory (Asher and Lascarides, 2003). While the usage of connectives can sometimes be optional, the *set* of connectives that a language offers is generally taken as important (if not exhaustive) evidence for the set of coherence relations that should be assumed.

1.1 Background: Connectives

From a syntactic viewpoint, ‘connective’ is not a homogeneous class, as it contains conjunctions, different kinds of adverbials, as well as certain prepositions. Our underlying definition of discourse connectives is based on (Pasch et al., 2003, p. 331):

(7) **Def.:** A *discourse connective* is a lexical item x that exhibits each of the following five properties:
(M1) x cannot be inflected.
(M2) x does not assign case features to its syntactic environment.
(M3) The meaning of x is a two-place relation.
(M4) The arguments of the relation (the meaning of x) are propositional structures.
(M5) The expressions of the arguments of the relation can be sentential structures.

Following (Stede, 2002), we drop M2 because our lexicon deliberately includes several prepositions that can be used as connectives (in the sense of M1, M3-M5), e.g., *trotz* (‘despite’) or *wegen* (‘due to’).

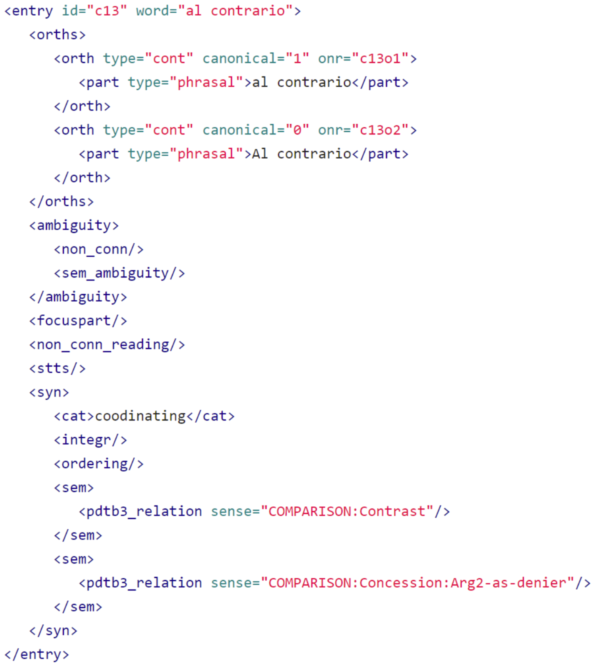
1.2 Motivation and contribution

Connectives can pose interesting challenges to translation and for language learners, as the differences in meaning between similar connectives can be quite subtle. For these reasons, we are interested here specifically in a *bilingual* Italian–German lexical resource, to be built on top of two existing single-language lexicons. As a case study, we focus on the subgroup of contrastive/concessive connectives, which we determined to comprise (in the existing lexicons) 31 German connectives and 12 Italian connectives; see Tables 3.2.2 and 3.2.2.

The main contributions of this paper are (1) suggestions for improving the existing language-specific resources used in this study through the technique of *cross-lingual projection* in a parallel corpus, which reveals correspondences between connectives and can point to gaps in either of the resources; and (2) an overview of the distribution of connectives and their senses, to be used in a bilingual database. Section 2 explains the two monolingual lexicons we work with, and Section 3 describes the corpus. Section 4 reviews related work in this area. Section 5 elaborates the idea of bilingual connective databases, and Section 6 summarises our findings.

2 Lexicons: DiMLex and LICo

We extracted the German contrastive connectives from DiMLex (Scheffler and Stede, 2016), a connective lexicon with several different fields describing orthographical variants, syntactic type, discourse sense, and usage examples. It contains 275 entries. The sense annotations are based on the Penn Discourse Treebank (PDTB) senses (Miltsakaki et al., 2008) in its latest version 3. The lexicon is publicly available[1] and aims to exhaustively describe the set of connectives for German, thus providing a basis for our case study.



```
<entry id="c13" word="al contrario">
   <orths>
      <orth type="cont" canonical="1" onr="c13o1">
         <part type="phrasal">al contrario</part>
      </orth>
      <orth type="cont" canonical="0" onr="c13o2">
         <part type="phrasal">Al contrario</part>
      </orth>
   </orths>
   <ambiguity>
      <non_conn/>
      <sem_ambiguity/>
   </ambiguity>
   <focuspart/>
   <non_conn_reading/>
   <stts/>
   <syn>
      <cat>coodinating</cat>
      <integr/>
      <ordering/>
      <sem>
         <pdtb3_relation sense="COMPARISON:Contrast"/>
      </sem>
      <sem>
         <pdtb3_relation sense="COMPARISON:Concession:Arg2-as-denier"/>
      </sem>
   </syn>
</entry>
```

Figure 1: *al contrario* entry in LICo

The set of Italian contrastive connectives comes from LICo (Feltracco et al., 2016), a similar lexicon for Italian containing 170 entries.[2] LICo was inspired by DiMLex and contains annotations on the same attributes and uses essentially the same structure (i.e., the same PDTB senses, orthographic variants, usage examples, etc.). An example entry of LICo is shown in Figure 1. We refer the reader to Feltracco et al. (2016) for details.

3 Exploiting a parallel corpus

For the parallel German/Italian corpus we used Europarl (Koehn, 2005), as it still appears to be the biggest resource of this kind, and it is, conveniently, already sentence-aligned. From the 1,832,053 sentences in the German-Italian part of the corpus we extracted the word alignments using MGIZA++ (Gao and Vogel, 2008). In the following, we sketch our method for obtaining the correspondence information on connectives based on these word alignments, and then present the results.

3.1 Method: Iterative lookup

We approach the problem from two sides: First we look up every German connective (31 in total) to get Italian alignments. 30 of them appeared in our Europarl corpus (with *dementgegen* missing). Then we look up every Italian connective to get German alignments (all 12 connectives present in

[1] https://github.com/discourse-lab/dimlex

[2] https://hlt-nlp.fbk.eu/technologies/lico

the corpus). We end up with a list of target language words or phrases (or empty elements, since a source language connective can also be covert in the target language) that are candidate contrastive connectives. Note that the lookup procedure does not differ structurally between words and phrases. In both cases, single words (stand-alone or in a phrase) can correspond to zero, one or more target words. The target representation is collected in a key-value structure, where the key is the position in the sentence and the value the word. This list is then sorted by position to return the target word or phrase (which is potentially discontinuous). Because the word alignment is not guaranteed to be correct, to filter for unlikely translations we focus on only the 3 most frequent alignments for every connective. We expect to find at least a subset of the already known (contrastive) connectives (from DiMLex or LICo), potentially complemented by a set of words or phrases that can help filling gaps in either of the lexicons.

This procedure produces at least some incorrect results for the following two reasons: 1) discourse connectives often can appear in a text with a connective reading or with a non-connective reading; and 2) connectives can have multiple senses, so that a connective may not have the contrastive reading in the particular sentence. The candidates produced hence have to be evaluated manually. Resulting candidates that have a connective reading are added to the seed list, in order to repeat the step back from the target language to the source language[3].

3.2 Results

3.2.1 German–Italian

The results of the first step of the iteration using the 31 German seed connectives are displayed in Table 3.2.2, where an underscore indicates an empty string (meaning that the connective was not aligned to a particular word or phrase in the target language) and the number after the underscore represents the (normalised) frequency of the alignment.

For the evaluation, we asked a native speaker of Italian with expert knowledge in linguistics to validate the resulting top 3 bilingual mappings. Firstly, we identified several possible connective candidates that were aligned to German contrastive connectives, but were not present in LICo, such as *al contempo, solo che, doppo tutto*. Secondly, we observed several possible orthographic variants of the already existing Italian connectives: *contro* or *contrario* (as possible variants of *al contrario*), and *d'altro canto* (as a variant of a discontinious connective *da un canto...dall'altro*). Finally, we found that several Italian connectives only had the *concession* sense, while the corresponding German connectives also had the Contrast sense, such as *comunque*, for which we found the German alignments *aber*, *allerdings* and *doch*, for example.

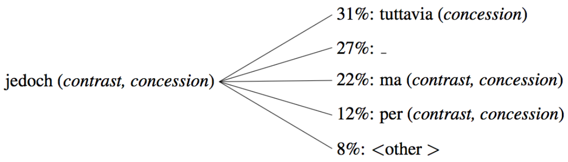

31%: tuttavia (*concession*)
27%: _
jedoch (*contrast, concession*)
22%: ma (*contrast, concession*)
12%: per (*contrast, concession*)
8%: <other >

Figure 2: Most frequent alignments of *jedoch*

As an example of a visualisation (for a single connective) the above analysis is based on, consider Figure 2, showing the most frequent alignments of *jedoch*, which always has a connective reading, thus nullifying the first problem mentioned in 3.1.

3.2.2 Italian–German

The results of the first step of the iteration using the 12 Italian seed connectives are displayed in Table 3.2.2. For 11 of the 12 contrastive connectives from LICo, the top 3 alignments yielded an existing DiMLex entry. The only connective without a DiMLex entry in the top 3 was *al contrario*, for which a possible new German connective candidate *im Gegenteil* was found through alignment.

Upon further investigation of the lower-ranked alignments (not included in Table 3.2.2), we were able to identify several other gaps in the German lexicon. Firstly, we observed that the Italian connective *invece* is frequently aligned to the German word *anstelle*, which is not in DiMLex (but *anstelle dessen* is). After examining the corresponding examples, we conclude that *anstelle* should be added to DimLex as a separate entry (similarly to the already existing *aufgrund* vs. *aufgrund dessen*). Also, we found that DiMLex lacks

[3] Ideally going back and forth until a stable and exhaustive set of candidates is found. For this study, we only did the first step, and then projected the found Italian connectives back to German.

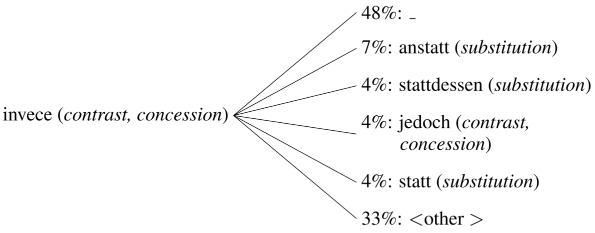

invece (*contrast, concession*)
48%: _
7%: anstatt (*substitution*)
4%: stattdessen (*substitution*)
4%: jedoch (*contrast, concession*)
4%: statt (*substitution*)
33%: <other >

Figure 3: Most frequent alignments of *invece*

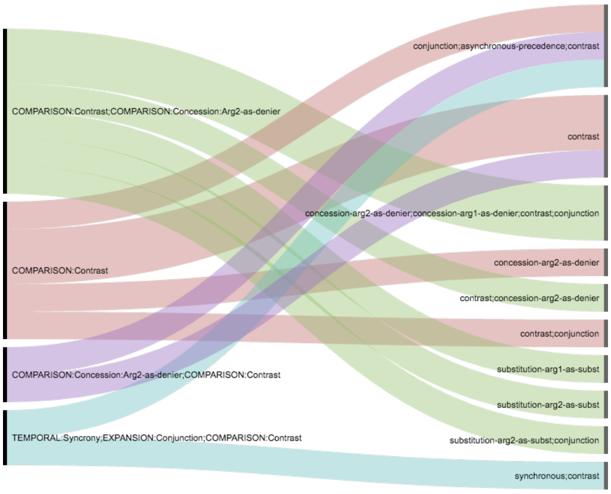


Figure 4: Mapping of connective senses from Italian to German

statt dessen as an orthographic variant of the more canonical *stattdessen*.

Finally, we identified two interesting cases that are DiMLex candidates: *umgekehrt* and *(ganz) im Gegenteil*, which we found aligned to the Italian *viceversa* and *al contrario*, respectively, but more corpus evidence is required to decide whether they can indeed serve as connective in the German language.

As an example visualisation, consider Figure 3, showing the most frequent alignments of *invece*, which always has a connective reading.

For Italian–German, we repeated the steps above with the candidates found using the German seed list (projecting the resulting Italian list back to German) to see if any additional connectives or orthographic variants would be found. We again found *im Gegenteil* through alignment of *al contrario* and a few alternative lexicalisations for DiMLex connectives[4], but no new candidates.

German connective (frequency)	Top 3 Italian alignments
aber (105413)	ma//_(0.24)//tuttavia
alldieweil (3)	finché//perché
allein (6973)	_(0.30)//solo//soltanto
allerdings (16232)	tuttavia//_(0.22)//ma
andererseits (6354)	_(0.30)//dall' altro//d' altro canto
bloß dass (117)	_(0.10)//solo che//che solo
dafür (36895)	_(0.70)//per//per aver
dafür // dass (42)	che//_(0.19)//per
dagegen (5423)	_(0.34)//contro//contrario
dahingegen (24)	_(0.17)//invece//al contrario
dementgegen (0)	
demgegenüber (121)	_(0.25)//invece//contro
doch (37423)	_(0.47)//ma//tuttavia
einerseits (4221)	da un lato//_(0.31)//da una parte
freilich (159)	_(0.30)//naturalmente//certo
gleichzeitig (13293)	_(0.35)//al contempo//allo stesso tempo
hingegen (1909)	invece//_(0.26)//tuttavia
immerhin (1360)	_(0.44)//comunque//dopo tutto
indessen (280)	invece//_(0.19)//tuttavia
jedoch (47667)	tuttavia//_(0.27)//ma
nur dass (21617)	che//solo che
sosehr (14)	malgrado tutto
unterdessen (193)	nel frattempo//_(0.21)//intanto
wiederum (2450)	_(0.55)//a sua volta//ancora una volta
wogegen (111)	mentre//_(0.19)//contro cosa
wohingegen (218)	mentre//_(0.14)//ma
während (20388)	_(0.28)//mentre//durante
währenddessen (78)	nel frattempo//_(0.17)//mentre
zugleich (3576)	_(0.41)//al contempo//allo stesso tempo
zum anderen (4299)	_(0.09)//altri//altre
zum einen (8848)	un//_(0.10)//una

Table 1: German connectives and their Italian alignments

Italian connective (frequency)	Top 3 German alignments
al contrario (3641)	im gegenteil//_(0.10)//im gegenteil
bensì (7107)	sondern//_(0.12)//sondern vielmehr
contrariamente a (661)	_(0.08)//entgegen//im gegensatz zu
da un canto (352)	einerseits//_(0.11)//andererseits
da un lato (4612)	einerseits//_(0.08)//einerseits die
da una parte (10194)	_(0.07)//und//eine
invece (18778)	_(0.48)//anstatt//stattdessen
ma (135218)	aber//sondern//_(0.15)
mentre (15773)	während//_(0.19)//und
per contro (13468)	gegen//und//_(0.06)
però (22687)	aber//jedoch//_(0.24)
viceversa (522)	umgekehrt//_(0.19)//hingegen

Table 2: Italian connectives and their German alignments

4 Related work

Parallel corpora have been successfully exploited before in order to automatically generate or induce connective lexicons in different languages. In particular, Versley (2010) projected discourse connectives across an English–German parallel corpus to train a discourse parser capable of disambiguating connective and non-connective readings. Similarly, Zhou et al. (2012) used an English–Chinese parallel corpus in order to build a Chinese connective lexicon via cross-lingual pro-

[4] Not listed here for reasons of space.

jection, and Hajlaoui and Popescu-Belis (2013) relied on parallel data to automatically retrieve Arabic counterparts for a subset of English connectives.

Since our goal was not to build a connective lexicon from scratch, but to extend the connective lists and refine the inventory of senses for the already existing lexicons, the closest approach to ours is the one adopted by Laali and Kosseim (2014), who aimed at automatically inducing a French connective lexicon via English–French parallel corpora using additional filtering rules. Similar to ours, their results have shown that using parallel translations can improve the coverage of the connective lists in both languages; however, since their lexicons used different sets of discourse relations, they were not able to extend their connective database in respect to senses, as opposed to our work.

5 Toward a bilingual connective database

Our study is meant as a step toward moving from single-language connective lexicons to a *bilingual* one that provides information about the relationships between the language-specific entries. Both monolingual lexicons are already publicly available on GitHub and in addition an interface allowing bilingual search has been made public in a related project[5]. Below we sketch additional plans for providing this information on the levels of connective tokens, and senses (coherence relations).

5.1 Connective mappings

One central purpose of a bilingual database is to assist translators (human or machine) or (human) language learners. For most connectives, there is a complicated m:n mapping between languages, which standard dictionaries do not cover, and the relevant features for making choices are not systematically known yet. A corpus-based inventory of mappings – ideally supplemented by pointers to the corpus instances and their context – can be a very useful resource for undertaking contrastive lexical investigations.

5.2 From connectives to phrases

The PDTB (Prasad et al., 2008) makes a distinction between connectives (a closed set) and “alternative lexicalizations” (AltLex), which are a non-demarcated set of phrases used to express a coherence relation. Such phrases are so far not part of DiMLex nor LICo. Obviously, they are much harder to detect: Corpus annotation (as done in PDTB) is one way, and we regard our cross-lingual projection method as another promising way. Quite often, connectives in language A have been translated to an AltLex in language B. We plan to study this more systematically by a closer inspection of the alignments and their contexts, in order to extract AltLex candidates as a supplement to the connective lexicons.

5.3 Senses and their distributions

A bilingual connective database can shed light on the distribution of senses over different languages and the degree of ambiguity that individual connectives exhibit. While we consider such conclusions premature for the current stage of the language-specific resources, we include Figure 4, which shows groups of connectives that share the same sense (or group of senses for ambiguous connectives) and their alignment to similar groups on the target side. The 12 Italian connectives (on the left), when grouped together based on their sense(s), form 4 sets, whereas for German (right side), fewer connectives (11 that were found in DiMLex among the top 3 alignments of the 12 source connectives) group into more sets (10). This suggests more ambiguity in Italian connectives, with less different senses represented by a larger set of connectives.

In addition, we observed that Italian connectives with a sense Contrast or Concession are frequently aligned to their German counterparts with a sense Substitution, such as *anstelle-invece.* Having examined the parallel examples more closely, we conclude that assigning both senses would be valid for both German and Italian, although they are placed distantly in the PDTB hierarchy of senses. These findings are confirmed by Feltracco et al. (2016), who acknowledge that the distinction between the two senses was one of the main cases of the inter-annotator disagreement. We conclude that both lexicons could benefit from adding additional senses gained via comparing parallel translations.

6 Summary

We present, to the best of our knowledge, the first Italian–German investigation of discourse connective lexicons. For the subclass of Contrast (in

[5] `http://connective-lex.info/`

a wide sense), we were able to identify several missing entries in both lexicons, and provided a start on identifying AltLex items for the two languages (future work). Once the information is organized in a complete bilingual database, it can assist translation and conclusions can be drawn regarding connective distribution, sense distribution and ambiguity in the different languages.

As prominent steps for future work, we note the disambiguation of connective- and non-connective readings, the implementation of more sophisticated filtering strategies to retrieve more reliable connective candidates and repeating this study for different languages pairs.

Acknowledgments

We are grateful to the Deutsche Forschungsgemeinschaft (DFG) for funding this work in the project 'Anaphoricity in Connectives'. We would like to thank the anonymous reviewers for comments on an earlier version of this manuscript. We also thank Flavia Adani for her help with translation and interpretation of the Italian results, and Tatjana Scheffler for the recent work on DiMLex.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Anna Feltracco, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. Lico: A lexicon of italian connectives. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, Italy.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *University of the Aegean-14th International Conference on Intelligent Text Processing and Computational Linguistics*. Springer.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *COLING*, pages 610–619.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi, 2008. *Sense annotation in the Penn Discourse Treebank*, pages 275–286. Springer Berlin Heidelberg, Berlin, Heidelberg.

Renate Pasch, Ursula Brauße, Eva Breindl, and UlrichH̃errmann Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Tatjana Scheffler and Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In Nicoletta Calzolari et al., editor, *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Manfred Stede. 2002. Dimlex: A lexical approach to discourse markers. In *Exploring the Lexicon - Theory and Computation*. Edizioni dell Orso, Alessandria.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. Northern European Association for Language Technology (NEALT).

Lanjun Zhou, Wei Gao, Bin Li, Zhong Wei, and Kam-Fai Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In *Proceedings of COLING*.